Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 98 (2007) 916–931

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

On the convergence of Newton's method when estimating higher dimensional parameters☆

Brenton R. Clarke[a], Andreas Futschik[b,*]

[a]*Mathematics and Statistics, Division of Science and Engineering, Murdoch University, Murdoch, Western Australia 6150, Australia*

[b]*Andreas Futschik, Department of Statistics, University of Vienna, Universitätsstr. 5/9, A-1010 Vienna, Austria*

Received 25 April 2005

Available online 30 December 2006

Abstract

In this paper, we consider the estimation of a parameter of interest where the estimator is one of the possibly several solutions of a set of nonlinear empirical equations. Since Newton's method is often used in such a setting to obtain a solution, it is important to know whether the so obtained iteration converges to the locally unique consistent root to the aforementioned parameter of interest. Under some conditions, we show that this is eventually the case when starting the iteration from within a ball about the true parameter whose size does not depend on n. Any preliminary almost surely consistent estimate will eventually lie in such a ball and therefore provides a suitable starting point for large enough n. As examples, we will apply our results in the context of M-estimates, kernel density estimates, as well as minimum distance estimates.

AMS 1991 subject classification: primary: 65U05secondary: 62H12

Keywords: Newton's algorithm; M-estimates; Kernel density estimates; Minimum distance estimates

1. Introduction

In many situations, estimates are obtained by solving a system of equations. If no closed form solution is available this has to be done numerically, with Newton's method being one of the most popular ways to obtain a solution. Like other methods, Newton's algorithm may not converge in

☆ Research support from a *Research Excellence Grant Scheme* Murdoch University, Murdoch, Western Australia 6150, Australia.

* Corresponding author.
E-mail addresses: B.Clarke@murdoch.edu.au (B.R. Clarke), Andreas.Futschik@univie.ac.at (A. Futschik).

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2006.12.002

some cases, and if it converges it may not converge against the desired solution. For instance, the gradient of the log-likelihood, or more generally an equation defining an M-estimate may have several zeros corresponding to different local extremes or even saddle points for example. One popular approach is thus to start Newton's iteration with a good initial estimate, hoping that this ensures convergence against the desired zero.

More technically, suppose that we want to estimate a parameter $\boldsymbol{\theta}_0$ defined as the solution of the equations $\boldsymbol{\lambda}(\boldsymbol{\theta}) = \mathbf{0}$ where $\boldsymbol{\lambda} : D \subseteq \mathbb{R}^r \to \mathbb{R}^r$. We assume the solution to be unique in some open ball B_0 of interest, but there may exist further solutions outside of B_0. Assume furthermore that a uniformly consistent estimate $\boldsymbol{\lambda}_n$ of $\boldsymbol{\lambda}$ is available at least on $\bar{B}_0$, the closure of B_0. Then a natural estimate of $\boldsymbol{\theta}_0$ is $\widehat{\boldsymbol{\theta}}_{0n}$ where $\widehat{\boldsymbol{\theta}}_{0n}$ satisfies $\boldsymbol{\lambda}_n(\widehat{\boldsymbol{\theta}}_{0n}) = \mathbf{0}$. Again $\widehat{\boldsymbol{\theta}}_{0n}$ is not necessarily unique. Frequently however, in such a setting, there is an asymptotically unique consistent root of the equations $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \mathbf{0}$. By this we mean that there is some ball B_δ of radius $\delta > 0$ about $\boldsymbol{\theta}_0$ where $\widehat{\boldsymbol{\theta}}_{0n}$ exists and is unique on B_δ for all sufficiently large n, and furthermore that $\widehat{\boldsymbol{\theta}}_{0n}$ converges to $\boldsymbol{\theta}_0$ almost surely. The $\widehat{\boldsymbol{\theta}}_{0n}$ solving

$$\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \mathbf{0} \tag{1}$$

may be hard or even impossible to obtain directly in higher dimensions. A common approach is thus to use Newton's algorithm starting from some initial estimate $\boldsymbol{y}_0$. For solving $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \mathbf{0}$, the algorithm is defined by the iteration

$$\boldsymbol{y}_{l+1}^{(n)} = \boldsymbol{y}_l^{(n)} - [\boldsymbol{\lambda}_n'(\boldsymbol{y}_l^{(n)})]^{-1}\boldsymbol{\lambda}_n(\boldsymbol{y}_l^{(n)}), \quad l = 0, 1, 2, \ldots . \tag{2}$$

We say that Newton's method converges, if there is an open ball B_δ centered at the true parameter $\boldsymbol{\theta}_0$ such that there is a unique solution $\widehat{\boldsymbol{\theta}}_{0n}$ of $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = 0$ in B_δ for all sufficiently large n, and furthermore Newton's method starting from any initial point x_0 in B_δ will converge to the desired estimate $\widehat{\boldsymbol{\theta}}_{0n}$ for all sufficiently large n. Since any strongly consistent preliminary estimate of $\boldsymbol{\theta}_0$, say $\hat{\boldsymbol{\theta}}_n^*$, will eventually be in B_δ, convergence of Newton's method applied to (1) with starting value $\hat{\boldsymbol{\theta}}_n^*$ implies that the Newton iteration will approach $\widehat{\boldsymbol{\theta}}_{0n}$ for all sufficiently large n.

Since Newton's algorithm may in general either not converge at all or converge to another than the desired zero, we intend to provide conditions under which a.s. convergence to $\widehat{\boldsymbol{\theta}}_{0n}$ holds when starting from some consistent initial estimate at least for large enough n. The reason why one may wish to use $\widehat{\boldsymbol{\theta}}_{0n}$ rather than say an initial consistent estimate is that the estimate $\widehat{\boldsymbol{\theta}}_{0n}$ may be more efficient (many examples of this are found in the literature). Indeed under suitable choices of $\boldsymbol{\lambda}_n$, a one step or even several step Newton iteration is frequently quoted as retaining the efficiency of the estimator based on (1).

It would be useful to know then if there was a fixed neighborhood based on the asymptotic curve $\boldsymbol{\lambda}(\boldsymbol{\theta})$ in which the Newton iteration would remain and indeed if the iteration was allowed to continue until fully iterated then one is assured that one will converge to the unique consistent root. Since it is not always the case that one has an initial consistent estimate, likely parameters are often based on physical phenomena and also plots of data (see [12]). One then iterates fully from likely initial estimates (perhaps several) and compares the solutions with the physical meaning of the problem. Another application of our result is in assessing empirically the performance of estimates through Monte Carlo simulation. Since one wants to know the empirical performance of the root $\widehat{\boldsymbol{\theta}}_{0n}$ for large but finite sample sizes which we know exists and is unique in the fixed neighborhood about $\boldsymbol{\theta}_0$ and there may not exist easily found initial consistent roots, then one can try starting from the hypothesized value $\boldsymbol{\theta}_0$ and for each sample generated fully iterating to the

desired unique consistent root in the fixed neighborhood. Example illustration of this approach in simulations are given in Clarke and McKinnon [15]. The current paper provides a justification for the authenticity of such empirical comparisons.

In the one-dimensional case and for M-estimates, the convergence of Newton's method has been investigated by Clarke [9]. In the current paper we consider the multi-dimensional case and discuss the required conditions in the context of three estimation problems that will be introduced in the subsequent three examples. Besides M-estimates, we consider both kernel density and minimum distance estimates as applications. Some smoothness assumptions are required to invoke our methodology. A discussion of these with relevance to M-estimates is given in Section 3.1.1, with regard to kernel density estimates in Section 3.2.1, and with regard to minimum distance estimates in Section 3.3.1. Without suitably smooth functions one may have to write down specific theory depending on the parametric or nonparametric model and the method of estimation, and Newton's method may well fail to converge (see Sections 3.1.1 and 3.2). With the aid of looking at smooth functions we have on the other hand a wide range of applications.

Example 1. Often M-estimates (or Z-estimates) are obtained by solving some equation system $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = 0$. Usually $\boldsymbol{\lambda}_n \to \boldsymbol{\lambda}$ pointwise or uniformly for some function $\boldsymbol{\lambda}$ satisfying $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = 0$. Robust M-estimates often require numerical methods for solving the system of equations.

Example 2. In nonparametric density estimation, the identification of modes is an important issue. The modes of a kernel density estimate are obvious estimates of the modes of the true density. Suppose we are looking for the mode in some region of interest. In higher dimensions, it can be very time consuming to locate a mode by evaluating some kernel density estimate on a very fine grid, in particular when there is a large number of data points. Carrying out the search on a not so fine grid, combined with an application of Newton's method starting from the grid point with the largest estimated density value is an obvious alternative.

Example 3. Minimum distance estimates are often obtained by looking for zeros of the gradient $\boldsymbol{\lambda}_n$ of some distance function. We will consider minimum Hellinger distance estimates, that are particularly useful in situations where a certain parametric model is assumed to be approximately, but not entirely correct. As shown by Beran [4], the estimates exhibit both desirable efficiency and robustness properties. The estimate needs to be calculated numerically and Beran [4] used Newton's method in his simulation study. (See [18] for a possible alternative algorithm.)

2. On the convergence of Newton's method

In this section, we first address the existence and uniqueness of the solution $\hat{\theta}_{0n}$ to $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \mathbf{0}$. We then provide conditions ensuring that Newton's method converges for all large enough n to this solution for any starting point in a sufficiently small ball about $\boldsymbol{\theta}_0$. The size of the ball does not depend on n.

More specifically, we assume that on the ball B_0 with center $\boldsymbol{\theta}_0$:

(A1) both $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_n$ are continuously differentiable,
(A2) all components and all partial derivatives of $\boldsymbol{\lambda}_n$ converge uniformly to those of $\boldsymbol{\lambda}$,
(A3) the Jacobian matrix $\boldsymbol{\lambda}'$ corresponding to the derivative of $\boldsymbol{\lambda}$ is continuous and has a nonzero determinant at $\boldsymbol{\theta}_0$.

Often but not always $\boldsymbol{\lambda}$ will be the gradient of some criterion function that is maximized or minimized. Think for instance of maximum likelihood estimates. In such a situation $\boldsymbol{\lambda}'$ actually is the Hessian matrix.

Lemma 1. *Assume that Assumptions* (A1)–(A3) *in this paper hold. Then for any sufficiently small ball B_δ with center θ_0 and radius δ, there is a unique root $\widehat{\boldsymbol{\theta}}_{0n}$ of $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \mathbf{0}$ in B_δ, for all sufficiently large n.*

Proof. The proof of the result is analogous to the proof in [22] for the maximum likelihood estimate. □

We now investigate the convergence of Newton's method when applied to (1). By $r(B)$, we denote the radius of a ball B. Furthermore, γ, resp. $\gamma^{(e)}$, will denote the—possibly different—Lipschitz constants required for $\boldsymbol{\lambda}$, resp. $\boldsymbol{\lambda}_n$.

Theorem 1. *Suppose that* (A1)–(A3) *are satisfied on B_0. Assume furthermore that*:

(N1) $\|\boldsymbol{\lambda}_n'(\boldsymbol{x})-\boldsymbol{\lambda}_n'(\boldsymbol{y})\| \leqslant \gamma^{(e)}\|\boldsymbol{x}-\boldsymbol{y}\|$ *for some $\gamma^{(e)}>0$, all $\boldsymbol{x}, \boldsymbol{y} \in B_0$ and for all sufficiently large n;*

(N2) $\|(\boldsymbol{\lambda}')^{-1}(\boldsymbol{x})\| \leqslant \beta$ *and* $\|(\boldsymbol{\lambda}')^{-1}(\boldsymbol{x})\boldsymbol{\lambda}(\boldsymbol{x})\| \leqslant \eta$ *for constants β and η satisfying $\alpha = \beta\eta\gamma^{(e)} < \frac{1}{2}$ on some ball $B_\delta \subset B_0$ satisfying $r(B_\delta)+t^* < r(B_0)$, where $t^* = [\beta\gamma^{(e)}]^{-1}[1-(1-2\alpha)^{1/2}]$.*

Then there exists a B_δ where δ can be chosen (based essentially on the asymptotic curve λ), to be small enough so that (N2) *holds. Furthermore, Newton's method applied to* (1) *with any starting value in B_δ will converge to $\hat{\theta}_{0n}$ for all sufficiently large n.*

Proof. According to Lemma 1, $\widehat{\boldsymbol{\theta}}_{0n} \in B_\delta$ and $\widehat{\boldsymbol{\theta}}_{0n}$ is the unique root in B_δ of $\boldsymbol{\lambda}_n$ for all sufficiently large n.

Let now $\beta^{(e)} = \beta + \epsilon$ and $\eta^{(e)} = \eta + \epsilon$ for some $\epsilon > 0$, chosen small enough such that $\beta^{(e)}\eta^{(e)}\gamma^{(e)} < \frac{1}{2}$, with $\gamma^{(e)}$ as in (N1). Due to (N2) and the uniform convergence of $\boldsymbol{\lambda}_n$ and $\boldsymbol{\lambda}_n'$, we have that both

$$\|(\boldsymbol{\lambda}_n')^{-1}(\boldsymbol{x})\| \leqslant \beta^{(e)} \quad \text{and} \quad \|(\boldsymbol{\lambda}_n')^{-1}(\boldsymbol{x})\boldsymbol{\lambda}(\boldsymbol{x})\| \leqslant \eta^{(e)} \quad \text{for all } x \in B_\delta,$$

f.a.s.l. n. The choice of B_δ according to (N2) ensures that the iteration will stay in B_0. Since all conditions of the Newton–Kantorovich Theorem applied to $\boldsymbol{\lambda}_n$ are satisfied, the result follows. □

Remark 1. Conditions (A1) and (A3) assure that there is a finite constant β such that $\|(\boldsymbol{\lambda}')^{-1}(\boldsymbol{x})\| \leqslant \beta$. Furthermore, since $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = \mathbf{0}$, it follows from the differentiability (and thus continuity) of $\boldsymbol{\lambda}$ that the condition $\|(\boldsymbol{\lambda}')^{-1}(\boldsymbol{x})\boldsymbol{\lambda}(\boldsymbol{x})\| \leqslant \eta$ holds for arbitrarily small $\eta > 0$ on a ball B_δ with sufficiently small radius δ. Thus, condition (N2) can be satisfied by choosing the radius δ of B_δ small enough.

As stated in Lemma 2 below, one way of assuring Lipschitz continuity of $\boldsymbol{\lambda}_n'$ (as required in condition (N1)) is to assume that both $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_n$ are twice continuously differentiable on $\bar{B}_0$ and also that the second partial derivatives of $\boldsymbol{\lambda}_n$ converge uniformly to those of $\boldsymbol{\lambda}$.

Lemma 2. *Assume that both $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}_n$ are twice continuously differentiable on $\bar{B}_0$ and that the second partial derivatives of $\boldsymbol{\lambda}_n$ converge uniformly to those of $\boldsymbol{\lambda}$. Then $\boldsymbol{\lambda}_n'$ is Lipschitz continuous on B_0.*

Let $h_{i,j}(\boldsymbol{x})$ denote the i, jth component of the matrix $\boldsymbol{\lambda}'(\boldsymbol{x})$. Then for all sufficiently large n, a Lipschitz constant $\gamma^{(e)}$ with respect to the maximum norm for $\boldsymbol{\lambda}_n'$ is given by $\gamma^{(e)} = \gamma + \epsilon$, where

$$\gamma = \max_{1 \leqslant i,j \leqslant k} \max_{z \in \bar{B}_0} \max_{l} |[\nabla h_{i,j}(z)]_l|$$

and $\epsilon > 0$ arbitrary.

Proof. We only consider the maximum norm. Since all vector and matrix norms on $\mathbb{R}^r$ are equivalent, Lipschitz continuity (possibly with another Lipschitz constant) follows for all other norms. By the mean value theorem and for some ξ on the line segment connecting $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$|h_{i,j}(\boldsymbol{x}) - h_{i,j}(\boldsymbol{y})| = |\nabla h_{i,j}(\xi)(\boldsymbol{x} - \boldsymbol{y})| \leqslant \max_{z \in \bar{B}_0} \|\nabla h_{i,j}(z)\| \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Thus,

$$\|\boldsymbol{\lambda}'(\boldsymbol{x}) - \boldsymbol{\lambda}'(\boldsymbol{y})\| \leqslant \gamma \|\boldsymbol{x} - \boldsymbol{y}\|.$$

Since $\bar{B}_0$ is compact, all maxima are finite. Analogous to $h_{i,j}$, define $h_{i,j}^{(n)}$ to be the i, jth component of $\boldsymbol{\lambda}_n'$. Due to uniform convergence of the second partial derivatives of $\boldsymbol{\lambda}_n$,

$$\|\nabla h_{i,j} - \nabla h_{i,j}^{(n)}\| \to 0 \quad \text{uniformly},$$

and Lipschitz continuity follows for $\boldsymbol{\lambda}_n'$. For any $\epsilon > 0$ and sufficiently large n, $\gamma^{(e)} = \gamma + \epsilon$ is a valid Lipschitz constant. □

3. Applications

In this section, we apply our Theorem 1 on the convergence of Newton's method to the situations of Examples 1–3. In particular, we will discuss when the required conditions (in particular, those concerning uniform convergence) will be satisfied.

3.1. M-Estimates

Given a set of independent identically distributed random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ on an observation space $\mathbb{R}^k$ with parametric distribution F_{θ_0}, many M-estimators can be written as a solution of a set of equations

$$\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}, \tag{3}$$

where we shall assume $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)$, so that each ψ_i has continuous partial derivatives $\frac{\partial}{\partial \theta_j} \psi_i(\boldsymbol{x}, \boldsymbol{\theta})$ which are again continuous in $\boldsymbol{x}$ and $\boldsymbol{\theta}$. See [8,11,15,16] for some particular applications. Under the usual assumption of Fisher consistency for M-estimators,

$$\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = \int \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta}_0) \, dF_{\theta_0}(\boldsymbol{x}) = 0,$$

for the true parameter $\boldsymbol{\theta}_0$.

We will now discuss conditions for our assumptions concerning the convergence of Newton's method to hold. Writing

$$\boldsymbol{\lambda}_n(\boldsymbol{\theta}) = \int \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})\, dF_n(\boldsymbol{x}),$$

where F_n is the empirical distribution function, the assumption of (A2) in fact can be shown from the theory of empirical processes. While pointwise strong consistency, that is

$$\boldsymbol{\lambda}_n(\boldsymbol{\theta}) \to \boldsymbol{\lambda}(\boldsymbol{\theta}) = \int \boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})\, dF_{\theta_0}(\boldsymbol{x}) \tag{4}$$

is a straightforward consequence of the strong law of large numbers the proof of uniform consistency needs some further arguments that will be sketched subsequently.

In robustness studies, it is usually the case that one considers a more general underlying distribution $G \in \mathcal{G}$ in a neighborhood of F_{θ_0} for the data (see [11] for example), but to illustrate the uniform convergence arguments, we consider only the case $G = F_{\theta_0}$ here.

To prove uniform convergence (Assumption (A2)), we will use Lemma 3 below taken from [7]. The result is similar to Theorem 23 in [24, p. 20]. Since we are interested in local arguments, it is sufficient to prove the result on some compact set D of parameter values $\boldsymbol{\theta}$.

Our result uses the notion of pointwise equicontinuity, which we define as in [33]: considering a function $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})$ on $\mathbb{R}^k \times D$, we define $\boldsymbol{\psi}$ to be "equicontinuous at each $\boldsymbol{x}$", if there is a neighborhood $N(\boldsymbol{x})$ for each $\boldsymbol{x}$ such that the class of functions $\mathcal{A} = \{\boldsymbol{\psi}_i(\cdot, \boldsymbol{\theta}) | \boldsymbol{\theta} \in D\}$ is equicontinuous on $N(\boldsymbol{x})$.

Lemma 3. *Let $\mathcal{A} = \{\boldsymbol{\psi}_i(\cdot, \boldsymbol{\theta}) | \boldsymbol{\theta} \in D\}$ be a family of real functions defined on $\mathbb{R}^k$ and suppose D is compact. Assume $\boldsymbol{\psi}$ is a continuous function in $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Then $\mathcal{A}$ forms an "equicontinuous at each $\boldsymbol{x}$" class of functions.*

A proof of this result is given in the appendix.

The following Theorem implies uniform convergence of $\boldsymbol{\lambda}_n$. Similar arguments can be used to establish uniform convergence for partial derivatives of $\boldsymbol{\lambda}_n$.

Theorem 2 (*Adapted from [33]*). *Let $\mathcal{A}$ be defined as above and $g(\boldsymbol{x})$ be a continuous function on $\mathbb{R}^k$ such that $\|\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta})\| < g(\boldsymbol{x})$ for each $\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{A}$ and $\boldsymbol{x} \in \mathbb{R}^k$. Suppose*

$$\int |g(\boldsymbol{x})|\, dF_{\theta_0}(\boldsymbol{x}) < \infty$$

then

$$\sup_{\theta \in D} |\boldsymbol{\lambda}_n(\boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{\theta})| \to 0. \quad a.s.$$

Remark 2. The above theorem can essentially be carried over to any ergodic sequence of random variables for which Eq. (4) holds. See [33].

Note also that Assumption (A3) of this paper is equivalent to assuming the matrix

$$\boldsymbol{M}(\boldsymbol{\theta}_0) = \int \left(\frac{\partial}{\partial \theta_j} \boldsymbol{\psi}_i(\boldsymbol{x}, \boldsymbol{\theta}) \right)_{(i,j)} \Bigg|_{\theta = \theta_0} dF_{\theta_0}(x)$$

is nonsingular, which is typical of M-estimation.

To illustrate the theory, we consider two examples.

Example 4. The maximum likelihood equations for the location and scale parameters $\boldsymbol{\theta} = (\mu, \sigma)$ of the parametric family of normal distributions on the real line lead to

$$\boldsymbol{\psi}(x; \mu, \sigma) = \left(\frac{x-\mu}{\sigma}, -1 + \left(\frac{x-\mu}{\sigma} \right)^2 \right)^\top.$$

Assuming $\sigma_0 > 0$ denote

$$D = \left\{ \begin{pmatrix} \mu \\ \sigma \end{pmatrix} \middle| \left\| \begin{pmatrix} \mu \\ \sigma \end{pmatrix} - \begin{pmatrix} \mu_0 \\ \sigma_0 \end{pmatrix} \right\| \leqslant \frac{\sigma_0}{2} \right\}.$$

Clearly $\boldsymbol{\psi}$ has continuous partial derivatives on D. (Thus (A1) is satisfied.) Since uniformly on D it is true that

$$\left\| \frac{x-\mu}{\sigma} \right\| < \left(\frac{2}{\sigma_0} \right) \left(|x-\mu_0| + \frac{\sigma_0}{2} \right),$$

the vector function $\boldsymbol{\psi}(x; \mu, \sigma)$ and the matrix of partial derivatives of $\boldsymbol{\psi}$ are bounded in Euclidean norm by

$$g(\boldsymbol{x}) = 1 + 4 \left(\frac{2}{\sigma_0} \right) \left(|x-\mu_0| + \frac{\sigma_0}{2} \right)^2 \max \left(1, \frac{\sigma_0}{2} \right),$$

which is clearly integrable with respect to the normal distribution with mean μ_0 and standard deviation σ_0. Assumption (A3) corresponds here to the nonsingularity of the Fisher Information matrix which is well known to be satisfied here. It is easily checked that (N1) is satisfied (one can either use similar arguments as to the above and show $\boldsymbol{\lambda}_n''$ converges uniformly to $\boldsymbol{\lambda}''$ and then implement Lemma 2 or else check directly). Condition (N2) follows from Remark 1.

The importance of this example is that even for nonrobust M-estimates, the Newton iteration starting from a consistent initial estimate of location and scale and applied to the M-estimating equations converges to the unique and consistent solution in this case which is the usual maximum likelihood estimator of location and scale.

Example 5. Consider the example of Clarke and McKinnon [15]. Here the statistical modeling and inference for a single ion channel is derived by using finite state space, continuous time Markov chains. The resulting dwell times in closed states for practical models discussed in that paper typically have a distribution

$$F_\theta(x) = \sum_{j=1}^{m} \epsilon_j G(x; \lambda_j) = \sum_{j=1}^{m} \epsilon_j (1 - e^{-\lambda_j x}), \tag{5}$$

where we have the constraints $\sum_{j=1}^{m} \epsilon_j = 1$, with (ϵ_j, λ_j) both positive, $j = 1, \ldots, m$; and $\lambda_i \neq \lambda_j$ for $i \neq j$. Estimators of the parameter θ are in that paper derived by minimizing

$$J_n(\theta) = \int_0^\infty (F_n(x) - F_\theta(x))^2 \, dx$$

and the resulting equations are shown to be solutions of M-estimating equations with $k = 2m-1$. It is easy to see from the appendices of that paper that the resulting ψ-function and its partial

derivatives with respect to θ are bounded in both x and θ, for $\theta \in D$, where D is a compact set away from the boundary of the parameter space. Hence, from Theorem 2, we have that, uniformly on D, all of $\boldsymbol{\lambda}_n$, $\boldsymbol{\lambda}_n'$ and $\boldsymbol{\lambda}_n''$ converge to $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}'$ and $\boldsymbol{\lambda}''$, respectively. Hence (A1) and (A2) are established. Assumption (A3) is shown to hold for this parametric family in the appendix by appealing to Lemma 3.1 of Clarke and Heathcote [14]. Thus, we can then appeal to Lemma 2 and then Remark 1 to establish (N1) and (N2).

The importance of this result is the following. The evaluation of the performance of that estimator is based on Monte Carlo simulations for data generated by the "generating parameter" $\boldsymbol{\theta}_0$, which is a solution of $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = 0$. It is typically the case in mixture estimation that there can be more than one zero to the estimating equations (3), which means that it can be important to converge to the unique consistent estimate $\widehat{\boldsymbol{\theta}}_{0n}$ (in a neighborhood of the true parameter $\boldsymbol{\theta}_0$) for large n, for all samples generated, when iterating until a root is found. It is important therefore that one can have reassurance that starting iterations from the true parameter $\boldsymbol{\theta}_0$ used for generating the data that one does not converge to some other root $\tilde{\boldsymbol{\theta}}_n$ that may be well away from $\boldsymbol{\theta}_0$. Just one or a few instances of such an occurrence can have an adverse effect on summary estimates either standard errors or mean squared errors for example. The fact that there is a ball or neighborhood of $\boldsymbol{\theta}_0$ independent of n for which starting at any point in the neighborhood would yield the unique consistent estimator $\widehat{\boldsymbol{\theta}}_{0n}$ in that neighborhood is a plus for evaluating the performance of the M-estimator in this situation by Monte Carlo methods and finding estimates of standard errors, at least for large sample sizes.

The results are not restricted to mixtures of exponential distributions. Similar calculations occur for estimators of parameters in finite mixtures of normal distributions, again evaluation of which has been carried out by Monte Carlo simulations starting from the generating parameter vector. (See [14].) Applications of these latter estimators in seismic data analysis are found in [12].

3.1.1. Further discussion regarding robust M-estimation

Robust M-estimators of a location parameter, $\theta_0 \equiv \mu_0$ given as the center of symmetry of an underlying distribution F defined on the real line, were introduced in [1,2,25,26,28]. Here one solves

$$\lambda_n(\mu) = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \mu) = 0.$$

The equations are frequently though not generally governed by ψ-functions that are continuous but only piecewise continuously differentiable. Exceptions for ψ being continuous are Huber's skipped mean, the median and the skipped median [27, p. 154] for example. Note here the root of the equations may not even exist or if it does it may not be an isolated root. The central distribution in all these discussions is $F = \Phi$, the standard normal distribution, though F may because of "contamination" be allowed to vary in small neighborhoods of Φ. Clarke [9] gives intimate details of the precise form of the domain of attraction of the M-estimator of location, and even gives limit distributions for the "boundary points" of the domain of attraction for these M-estimators. That is, one can go beyond finding just existence of a ball B_δ where for all sufficiently large n starting from any point in B_δ Newton's method converges to the consistent root of the equations. However, the form of ψ needs to be explicitly given and even then, arguments rely initially on the local convergence result, that is, where δ is chosen small enough and based on the asymptotic equation. Then the precise nature of ψ and $F = \Phi$ is used to determine explicitly a wider region from which the Newton iteration will converge into B_δ.

Hampel et al. [27] extend the ideas of M-estimation to fairly general parametric models using optimal "B-robustness" and optimal "V-robustness". Again these methods often yield a choice of $\boldsymbol{\psi}$-function that is continuous though only piecewise continuously differentiable. Arguments using Frank Clarke's [17] book on nonsmooth analysis can potentially be employed in order to attempt to establish existence of the regions B_δ so that for all sufficiently large n starting from anywhere in B_δ the Newton iteration will converge to the consistent root $\widehat{\boldsymbol{\theta}}_{0n}$. In this case it much depends on the particular $\boldsymbol{\psi}$, the particular parametric model $F_{\boldsymbol{\theta}}$ and indeed the particular parameter $\boldsymbol{\theta}_0$ as well as the observation space ($\mathbb{R}$ or $\mathbb{R}^k$). Example application of nonsmooth analysis for proving Fréchet differentiability of resulting statistical functionals is given in [10]. Uniform convergence arguments for the partial derivatives of $\boldsymbol{\lambda}_n(\boldsymbol{\theta})$ which in this case involve generalized Jacobians (see [10, Definition 2.1]) to

$$\boldsymbol{\lambda}'(\boldsymbol{\theta}) = E_{F_{\theta_0}}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\psi}(\boldsymbol{X},\boldsymbol{\theta})\right]$$

rely here on uniform convergence over classes of functions and classes of sets. This theory is explained in that paper only for observations on the real line, and sets of the form $(-\infty, x]$. On the other hand, Hampel et al. [27] consider parametric models for more general observation spaces such as $\mathbb{R}^k$, $k > 1$. The methodology appears complicated, depends on the choice of $\boldsymbol{\psi}$ and F_{θ_0}. All this can be obviated by choosing $\boldsymbol{\psi}$ to be smooth. Bednarski and Zontek [3] illustrate how one can avoid the curse of $\boldsymbol{\psi}$-functions with sharp corners, even in a quite complicated parametric model.

3.2. Kernel density estimation

Consider a three times continuously differentiable density function f on $\mathbb{R}^k$ having a mode at $\boldsymbol{\theta}_0 \in B_0$. We assume this mode to be unique in B_0. Then obviously $\boldsymbol{\theta}_0$ solves $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = \mathbf{0}$, with $\boldsymbol{\lambda}$ being the gradient of f. Here $\boldsymbol{\lambda} : \mathbb{R}^k \to \mathbb{R}^k$, so $k = r$ in this case. Let us similarly denote $\boldsymbol{\lambda}_n$ the gradient of the kernel estimate

$$\hat{f}_n(\boldsymbol{t}) := \frac{1}{nh_n^k}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{t}-\boldsymbol{X}_i}{h_n}\right). \tag{6}$$

An estimate of the mode is obtained by solving $\boldsymbol{\lambda}_n(\boldsymbol{t}) = \mathbf{0}$.

We now state a set of conditions ensuring convergence of the Newton estimate to the desired mode $\boldsymbol{\theta}_0$, when starting the iteration from a point sufficiently close to $\boldsymbol{\theta}_0$. Besides sufficient smoothness of the underlying density f, we need that f is not too flat at the mode, i.e. that the Hessian matrix of second partial derivatives at $\boldsymbol{\theta}_0$ is negative definite. Sufficient smoothness of $\hat{f}_n(\boldsymbol{t})$ can be ensured by choosing an at least three times continuously differentiable kernel satisfying condition (K1) of [23], like the biweight or the Gaussian kernel. It is interesting to note that for less smooth kernel functions, like the uniform or the Epanechnikov kernel, Newton's method will not converge in general. Indeed even in the univariate case, Newton's method is not well defined for the uniform kernel $K(x) = \frac{1}{2}\mathbb{1}_{[-1\leqslant x\leqslant 1]}$. For the Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{[-1\leqslant x\leqslant 1]}$, it is easy to see that a Newton step always leads to the average of those observations that are within a distance of h_n from the previous estimate and thus Newton's method cannot be expected to converge in this setting. On the other hand, simulations indicate that Newton's method seems to work for the smoother (but still not three times continuously differentiable) biweight kernel $K(x) = \frac{15}{16}(1-x^2)^2\mathbb{1}_{[-1\leqslant x\leqslant 1]}$.

Under sufficient smoothness, consistency can be ensured by choosing the bandwidth of the kernel estimate in the range $cn^{-\alpha_1} < h_n < cn^{-\alpha_2}$ for some $c > 0$ and $0 < \alpha_2 < \alpha_1 < \frac{1}{k+m}$. A more detailed discussion can be found in Section 3.2.1.

3.2.1. Convergence conditions

Recall that by Lemma 1, conditions (A1)–(A3) ensure a unique solution to $\boldsymbol{\lambda}_n(\boldsymbol{t}) = \mathbf{0}$ in some ball B_0 about the true mode $\boldsymbol{\theta}_0$. For $\boldsymbol{\lambda}$, (A1) and (A3) (and also (N2)) will be trivially satisfied, if the Hessian matrix of second partial derivatives at $\boldsymbol{\theta}_0$ is negative definite.

For $\boldsymbol{\lambda}_n$, condition (A1) can be met by choosing a sufficiently smooth kernel function, for instance the Gaussian kernel. We will now have a closer look at condition (A2) concerning uniform convergence in this context. In the one-dimensional case, the uniform convergence of the kernel density estimate and its derivatives has been explored in different settings by several authors. A classical paper is by Silverman [34], for further work in this direction see for instance Devroye and Wagner [20], or Karunamuni and Mehra [29]. Results for the higher dimensional case have been obtained, among others, by Deheuvels [19] and Bertrand-Retali [5]. To establish (A2), a recent paper addressing the higher dimensional case by Giné and Guillou [23] will be useful in our context. To establish uniform convergence, it suffices to show componentwise uniformity. For this, assume λ^* to be either a component of λ or a first or second order partial derivative of a component of λ. Define an analogous generic component for λ_n^*. Notice for this purpose that the natural kernel estimate of some mth order partial derivative is given by the respective derivative of the kernel estimate, i.e. $\frac{\partial}{\partial t_{i_1} \cdots \partial t_{i_m}} f(\boldsymbol{t})$ is estimated by

$$\frac{1}{n h_n^{k+m}} \sum_{i=1}^{n} K^{(i_1,\ldots,i_m)} \left(\frac{\boldsymbol{t} - \boldsymbol{X}_i}{h_n} \right),$$

with $K^{(i_1,\ldots,i_m)}(\boldsymbol{u})$ denoting the partial derivative $\frac{\partial}{\partial t_{i_1} \cdots \partial t_{i_m}} K(\boldsymbol{u})$.

Since

$$\sup_{\boldsymbol{t} \in B_0} |\lambda_n^*(\boldsymbol{t}) - \lambda^*(\boldsymbol{t})| \leqslant \sup_{\boldsymbol{t} \in B_0} |\lambda_n^*(\boldsymbol{t}) - E\lambda_n^*(\boldsymbol{t})| + \sup_{\boldsymbol{t} \in B_0} |E\lambda_n^*(\boldsymbol{t}) - \lambda^*(\boldsymbol{t})|,$$

the bias and the stochastic component may be addressed separately. To establish (A2), uniform convergence of all partial derivatives up to order $m = 2$ of $\hat{f}_n$ to those of f needs to be shown. Furthermore, according to Lemma 2, uniform convergence of all third partial derivatives is sufficient to establish (N1). For the stochastic part, [23, Section 2] provides conditions leading to uniform convergence of the kernel density estimate and all partial derivatives of the required order to its expected value. We comment briefly on their requirements. Their condition (K1) permits for quite general sufficiently smooth kernel functions. Indeed when applied to applied to $K^{(i_1,\ldots,i_m)}(\boldsymbol{u})$ it is satisfied for instance for the Gaussian kernel, i.e. K being any (i.i.d. or more general) multivariate normal density function. Their bandwidth conditions translate into $h_n \to 0$, $\frac{n h_n^{k+m}}{|\log h_n|} \to \infty$, $\frac{|\log h_n|}{\log \log n} \to \infty$ as $n \to \infty$. Furthermore, we need to assume that there is a constant c such that $h_n^{k+m} \leqslant c h_{2n}^{k+m}$. For the underlying density only boundedness is needed for the stochastic part.

The bias part

$$\sup_{t \in B_0} |\mathrm{E}(\hat{f}_n^{(i_1,\ldots,i_m)}(\boldsymbol{t}) - f^{(i_1,\ldots,i_m)}(\boldsymbol{t}))|$$

can be dealt with by standard arguments like those given by Silverman [34] for the one-dimensional case. To obtain, for instance, uniform convergence of the bias of some second partial derivative it is sufficient that $h_n \to 0$ and both the underlying density f and the kernel K used are three times continuously differentiable.

To summarize, conditions (A2) and (N1) will be satisfied when the underlying density is sufficiently smooth, the kernel function K in (6) is chosen to be at least three times continuously differentiable, and a suitable bandwidth (for instance of the MSE-optimal rate $h_n = cn^{-1/(4+k)}$) is chosen.

3.3. Minimum distance estimates

The minimum distance principle often leads to estimates and hypothesis tests exhibiting desirable properties. Here, we focus on minimum Hellinger distance estimates, as proposed by Beran [4]. Given a parametric family of densities $\mathcal{F} = \{f_\theta : \boldsymbol{\theta} \in D\}$, the estimate is defined as a $\widehat{\boldsymbol{\theta}}_{0n}$ minimizing

$$\| f_\theta^{1/2} - \hat{f}_n^{1/2} \|^2 \tag{7}$$

in $\boldsymbol{\theta}$, where $\hat{f}_n$ is a nonparametric estimate of the underlying density f, for instance the kernel density estimate introduced in (6). Notice that minimizing (7) is equivalent to maximizing the empirical affinity

$$\int f_\theta^{1/2}(\boldsymbol{x}) \hat{f}_n^{1/2}(\boldsymbol{x})\, d\boldsymbol{x} \tag{8}$$

with respect to $\boldsymbol{\theta}$. For this purpose one commonly solves

$$\boldsymbol{\lambda}_n(\boldsymbol{\theta}) := \int \frac{\partial}{\partial \boldsymbol{\theta}} f_\theta^{1/2}(\boldsymbol{x}) \hat{f}_n^{1/2}(\boldsymbol{x})\, d\boldsymbol{x} = \mathbf{0}.$$

The goal is then to converge to the "correct" parameter value $\boldsymbol{\theta}_0$ maximizing the theoretical affinity

$$\int f_\theta^{1/2}(\boldsymbol{x}) f^{1/2}(\boldsymbol{x})\, d\boldsymbol{x} \tag{9}$$

with f denoting the true density.

One way to obtain an estimate is to search for a zero of the gradient of (7) w.r.t. $\boldsymbol{\theta}$ via Newton's method. As pointed out by Beran [4], the solution to (7) may not always be unique, and he therefore proposes to start Newton's method from an initial robust parameter estimate. As in the previous subsections, conditions need to be satisfied for Newton's method to converge against a correct solution and thus to provide consistent estimates. Again, we briefly discuss how the conditions of our Theorem 1 can be established. Requirement (A1) can be satisfied by assuming that $f_\theta^{1/2}$ is sufficiently smooth, in particular the existence of the first three derivatives, with all third order partial derivatives being continuous with a finite integral. Similar assumptions have been made by Beran [4]. If the true density satisfies $f \in \mathcal{F}$, (A3) is an assumption to be checked for the specific considered parametric family $\mathcal{F}$. If (A3) holds for $f \in \mathcal{F}$, it still can be expected to hold also for sufficiently small deviations from the parametric model.

For (N2), see Remark 1. To establish (A2) and (N1), it is sufficient to show uniform (in $\boldsymbol{\theta}$) convergence of all partial derivatives up to order three of (8) to those of (9). If (according to our assumption) these partial derivatives have finite integrals, uniform convergence follows under

the conditions ensuring the almost sure uniform convergence of the kernel density estimate as discussed in Section 3.2. Since the partial derivatives are with respect to $\boldsymbol{\theta}$ and not $\boldsymbol{x}$, derivatives of the kernel estimate need not to be considered. Thus, in contrast to Section 3.2, the degree of smoothness of the kernel function is not essential here. We give a brief heuristic justification for the first partial derivatives occurring in $\boldsymbol{\lambda}_n$, under the assumption that $\int \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x})\,d\boldsymbol{x}$ is finite. Consider for this purpose an arbitrary component θ_i of $\boldsymbol{\theta}$. Taking an arbitrarily small $\epsilon > 0$, uniform convergence can be shown by establishing ϵ as an a.s. uniform bound on the difference between the respective partial derivatives of the theoretical (9) and the empirical affinities (8) for sufficiently large $n = n(\epsilon, \omega)$.

$$
\begin{aligned}
&\left| \int \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x}) \hat{f}_n^{1/2}(\boldsymbol{x})\,d\boldsymbol{x} - \int \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x}) f^{1/2}(\boldsymbol{x})\,d\boldsymbol{x} \right| \\
&= \left| \int \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x}) \left[\left(f(\boldsymbol{x}) - (f(\boldsymbol{x}) - \hat{f}_n(\boldsymbol{x})) \right)^{1/2} - f(\boldsymbol{x})^{1/2} \right] d\boldsymbol{x} \right| \\
&\leqslant \left| \int \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x}) \left[\left(f(\boldsymbol{x}) + |f(\boldsymbol{x}) - \hat{f}_n(\boldsymbol{x})| \right)^{1/2} - f(\boldsymbol{x})^{1/2} \right] d\boldsymbol{x} \right| \\
&\leqslant \left| \int_{x: f > \epsilon^2} \right| + \left| \int_{x: f \leqslant \epsilon^2} \right| \qquad (10) \\
&\leqslant \left| \int_{x: f > \epsilon^2} \frac{\partial}{\partial\theta_i} f_\theta^{1/2}(\boldsymbol{x}) \left| \frac{\hat{f}_n(\boldsymbol{x}) - f(\boldsymbol{x})}{\epsilon} \right| d\boldsymbol{x} \right| + c\epsilon \qquad (11) \\
&\leqslant c\epsilon \qquad (12)
\end{aligned}
$$

for all large enough $n = n(\epsilon, \omega)$ such that $\|\hat{f}_n - f\|_\infty < \epsilon^2$, and for some generic constants c. Inequality (11) follows by a first order Taylor expansion of $\sqrt{y_0 + y} - \sqrt{y_0}$ at $y_0 = f(\boldsymbol{x})$ and for $y = |f(\boldsymbol{x}) - \hat{f}_n(\boldsymbol{x})|$, while observing that the remainder term is negative.

3.3.1. Minimum distance estimates more generally

In 1970s and 1980s numerous contributions to minimum distance estimation were made. We have discussed the L_2-minimum distance estimator given in Example 5 which is also an M-estimator, and the Hellinger distance estimator of [4,18] in Example 3. There are many more minimum distance based methods. See an early comment by Clarke and Heathcote [13] on the paper by Quandt and Ramsey [32]. Another example includes the Cramér von Mises distance estimator of [30] and the paper by Woodward et al. [35]. All the distance measures mentioned here involve quantities realizing uniform convergence of the $\boldsymbol{\lambda}_n(\boldsymbol{\theta}) \to \boldsymbol{\lambda}(\boldsymbol{\theta})$ either through uniform convergence say of the empirical characteristic function to the characteristic function of the parametric distribution, or the empirical moment generating function to the moment generating function of the parametric distribution, or of the empirical distribution function to the parametric distribution, respectively, say. Similar results hold for the partial derivatives of $\boldsymbol{\lambda}$ though it would become tedious to give all details here.

However, we do *not* claim all distance based methods yield regions B_δ such that Newton's method converges. One need only look no further than the L_1-distance estimator for which there is ample discussion of suitable algorithms to evaluate estimators. See [6,21].

Table 1
Proportion of cases where Newton's method does not converge to the root of the empirical equation system that is closest to the true parameter values

Initial value	Normal data			Contaminated normal data		
	$n = 20$	$n = 40$	$n = 100$	$n = 20$	$n = 40$	$n = 100$
Robust	0.002	0	0	0.030	0.002	0
Nonrobust	0.006	0	0	0.182	0.170	0.054

4. Simulation

The following small simulation study is intended to illustrate the issues raised in the paper from a practical point of view. We investigate the convergence of Newton's method when applied to minimum distance estimates of the parameter vector (μ, σ) of the normal distribution. In the setting presented in [4, Section 6], we consider three different sample sizes ($n = 20, 40, 100$), and two different choices of the starting values for the Newton iteration. As nonrobust starting values, we choose the sample mean $\bar{x}$ and the sample standard deviation s. On the other hand, the sample median $\tilde{x}$ and the median absolute deviation $(0.674)^{-1}\text{median}|x_i - \tilde{x}|$ are taken as robust alternatives. Based on Newton's method we try to identify the parameter values (μ, σ) that minimizes the Hellinger distance between the respective normal density and a kernel density estimate based on the Epanechnikov kernel and with bandwidths $h_n = 0.7\left(\frac{40}{n}\right)^{1/5}$ chosen analogous to the recommendation of [4]. For this purpose and for each setting, 500 samples were generated either from a normal $N(0, 1)$ distribution ("normal data") or from the gross error model $0.9N(0, 1) + 0.1N(0, 3^2)$ ("contaminated normal data").

Table 1 displays the proportion of cases where Newton's method fails to converge to the root closest to the true parameter. In accordance with our theoretical results, the performance of Newton's method is good, when the initial estimates are not too far away from the true parameter values. This is obvious when looking at the nonrobust starting values in the contaminated model. Furthermore, the performance improves with increasing sample size.

Appendix A.

Our Theorem 1 uses the below result concerning the convergence of Newton's method which can be found in [31, Section 12.6.2]. By $\|\cdot\|$, we denote either some vector norm on $\mathbb{R}^r$ or the corresponding matrix norm for $r \times r$ matrices.

Theorem 3 (*Newton–Kantorovich Theorem*). *Assume that $\boldsymbol{\lambda} : D \subset \mathbb{R}^r \to \mathbb{R}^r$ is differentiable on a convex set D_0 and that*

$$\|\boldsymbol{\lambda}'(\boldsymbol{x}) - \boldsymbol{\lambda}'(\boldsymbol{y})\| \leqslant \gamma\|\boldsymbol{x} - \boldsymbol{y}\| \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in D_0.$$

Suppose that there is an x_0 in D_0 such that $\|\boldsymbol{\lambda}'(\boldsymbol{x}_0)^{-1}\| \leqslant \beta$ and $\alpha = \beta\gamma\eta \leqslant \frac{1}{2}$ where $\eta \geqslant \|\boldsymbol{\lambda}'(\boldsymbol{x}_0)^{-1}\boldsymbol{\lambda}(\boldsymbol{x}_0)\|$. Set

$$t^* = (\beta\gamma)^{-1}[1 - (1 - 2\alpha)^{1/2}],$$

and let $\bar{B}(\boldsymbol{x}_0, t^*)$ denote a closed ball with center x_0 and radius t^*. If $\bar{B}(\boldsymbol{x}_0, t^*) \subset D_0$, then the Newton iterates are well defined, remain in $\bar{B}(\boldsymbol{x}_0, t^*)$ and converge to a solution x^* of $\boldsymbol{\lambda}(\boldsymbol{x}) = 0$.

Proof (*of Lemma 3*). Take any fixed $\boldsymbol{x}$, and let $C(\boldsymbol{x})$ be some compact set containing a neighborhood of $\boldsymbol{x}$. Since D is compact, $\boldsymbol{\psi}$ is uniformly continuous on $C(\boldsymbol{x}) \times D$.

Thus, for all $\epsilon > 0$ and $\boldsymbol{y} \in C(\boldsymbol{x})$ there is a δ such that

$$\|\boldsymbol{\psi}(\boldsymbol{x}, \boldsymbol{\theta}) - \boldsymbol{\psi}(\boldsymbol{y}, \boldsymbol{\theta})\| < \epsilon,$$

whenever

$$\left\| \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{\theta} \end{pmatrix} \right\| < \delta.$$

Equicontinuity at x now follows since

$$\left\| \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\theta} \end{pmatrix} - \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{\theta} \end{pmatrix} \right\| = \|\boldsymbol{x} - \boldsymbol{y}\|. \qquad \square$$

Appendix B.

Condition (A3) in relation to Example 5 is such that one wishes to show the matrix $\boldsymbol{M}(\boldsymbol{\theta}_0)$ is nonsingular. From Clarke and McKinnon [15] this is true whenever the matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}[(\partial^2/\partial\boldsymbol{\theta}^2)\frac{1}{2}J_n(\boldsymbol{\theta})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is nonsingular. The latter matrix has as its elements

$$\lambda_{i,j}(\boldsymbol{\theta}) = \int_{\mathbb{R}^+} \{(\partial/\partial\theta_i)F_{\boldsymbol{\theta}}(y)\}\{(\partial/\partial\theta_j)F_{\boldsymbol{\theta}}(y)\}\,dy.$$

We now repeat Lemma 3.1 of [14] adapted here to the parametric model of exponential mixtures on the positive real line whereupon there are $2m-1$ free parameters and illustrate in the example of a nondegenerate mixture of two exponential distributions that the matrix $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ is nonsingular.

Lemma 4. *Assume that there does not exist a nonzero vector* $\boldsymbol{b}^\top = (b_1, \ldots, b_{2m-1})$ *such that*

$$\boldsymbol{b}^\top(\partial/\partial\boldsymbol{\theta})F_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{2m-1} b_i(\partial/\partial\theta_i)F_{\boldsymbol{\theta}}(x) = 0 \quad \textit{for every } x \in (0,\infty). \tag{13}$$

Then for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, *the matrix* $\boldsymbol{\Lambda}(\boldsymbol{\theta})$ *and consequently the matrix* $\boldsymbol{M}(\boldsymbol{\theta})$ *is nonsingular when* $\boldsymbol{\psi}$ *is given as in* [15, Appendix].

Proof. As an illustration of how (13) cannot be satisfied for any nonzero $\boldsymbol{b}$, we consider $m = 2$. Here then $\boldsymbol{\theta} = (\epsilon, \lambda_1, \lambda_2)$, whereupon $\epsilon \equiv \epsilon_1$ and $(1-\epsilon) \equiv \epsilon_2$ in the model (5). Then

$$b_1(e^{-\lambda_2 x} - e^{-\lambda_1 x}) + b_2\epsilon x e^{-\lambda_1 x} + b_3(1-\epsilon)x e^{-\lambda_2 x} = 0$$

or equivalently

$$q(x, \boldsymbol{b}, \boldsymbol{\theta}) = b_1(e^{(\lambda_1-\lambda_2)x} - 1) + b_2\epsilon x + b_3(1-\epsilon)x e^{(\lambda_1-\lambda_2)x} = 0.$$

Now

$$\begin{aligned} q'(x, \boldsymbol{b}, \boldsymbol{\theta}) = {} & (\lambda_1 - \lambda_2)b_1 e^{(\lambda_1-\lambda_2)x} + b_2\epsilon \\ & + b_3(1-\epsilon)e^{(\lambda_1-\lambda_2)x} + b_3(1-\epsilon)(\lambda_1-\lambda_2)x e^{(\lambda_1-\lambda_2)x} = 0. \end{aligned}$$

So $q''(x, \boldsymbol{b}, \boldsymbol{\theta})$ equals

$$\{(\lambda_1-\lambda_2)^2 b_1 + b_3(1-\epsilon)(\lambda_1-\lambda_2) + b_3(1-\epsilon)(\lambda_1-\lambda_2) + b_3(1-\epsilon)(\lambda_1-\lambda_2)^2 x\} e^{(\lambda_1-\lambda_2)} = 0.$$

This implies in particular that the quantity in braces immediately above is zero uniformly in x. Subsequently, b_3 must be zero which further implies $b_1 = 0$ whence $b_2 = 0$. That is, there does not exist a nonzero $\boldsymbol{b}$ such that $\boldsymbol{b}^\top(\partial/\partial\boldsymbol{\theta})F_{\boldsymbol{\theta}}(x) = 0$ uniformly in x in the two component case. Therefore, in the case of a mixture of two component exponentials with components having differing means it can be concluded from Lemma 4 that the matrix $\boldsymbol{M}(\boldsymbol{\theta})$ is nonsingular. In fact $\boldsymbol{M}(\boldsymbol{\theta})$ is positive definite.

More generally in the case of this L_2 estimator for the model $F_{\boldsymbol{\theta}}$, for instance as in (5), the matrix $\boldsymbol{M}(\boldsymbol{\theta})$ can be shown to be nonsingular if the functions given by the partial derivatives of $F_{\boldsymbol{\theta}}$ are linearly independent. This involves checking the Wronskian and showing it is nonzero for at least one point x. □

References

[1] D.F. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers, J.W. Tukey, Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, NJ, 1972.

[2] A.E. Beaton, J.W. Tukey, The fitting of power series meaning polynomials, illustrated on band-spectroscopic data, Technometrics 16 (1974) 147–185.

[3] T. Bednarski, S. Zontek, Robust estimation of parameters in a mixed unbalanced model, Ann. Statist. 24 (1996) 1493–1510.

[4] R. Beran, Minimum Hellinger distance estimates for parametric models, Ann. Statist. 5 (1977) 445–463.

[5] M. Bertrand-Retali, Convergence uniforme d'un estimateur de la densité par la méthode du noyau, Rev. Roumaine Math. Pures Appl. 23 (1978) 361–385.

[6] P. Bloomfield, W.L. Steiger, Least Absolute Deviations, Theory Applications and Algorithms, Birkhäuser, Basel, 1983.

[7] B.R. Clarke, Robust estimation, limit theorems and their applications, Ph.D. Thesis, Australian National University, 1980.

[8] B.R. Clarke, Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations, Ann. Statist. 11 (1983) 1196–1205.

[9] B.R. Clarke, Asymptotic theory for description of regions in which Newton–Raphson iterations converge to location M-estimators, J. Statist. Plann. Inference 15 (1986) 71–85.

[10] B.R. Clarke, Nonsmooth analysis and Fréchet differentiability of M-functionals, Probab. Theory Related Fields 73 (1986) 197–209.

[11] B.R. Clarke, A remark on robustness and weak continuity of M-estimators, J. Austral. Math. Soc. Ser. A 68 (2000) 411–418.

[12] B.R. Clarke, A review of differentiability in relation to robustness with application to seismic data analysis, PINSA Ser. A 66 (2000) 467–482.

[13] B.R. Clarke, C.R. Heathcote, Comment on "Estimating mixtures of normal distributions and switching regressions", J. Amer. Statist. Assoc. 73 (1978) 749–750.

[14] B.R. Clarke, C.R. Heathcote, Robust estimation of k-component univariate normal mixtures, Ann. Inst. Statist. Math. 46 (1994) 83–93.

[15] B.R. Clarke, P.L. McKinnon, Robust inference and modeling for the single ion channel, J. Statist. Comput. Simulation 75 (2005) 513–529.

[16] B.R. Clarke, G.F. Yeo, R.K. Milne, Local asymptotic theory for multiple solutions of likelihood equations, with application to a single ion channel model, Scand. J. Statist. 20 (1993) 133–146.

[17] F.H. Clarke, Optimization and Nonsmooth Analysis, Wiley, New York, 1983.

[18] A. Cutler, O.I. Cordero-Braña, Minimum Hellinger distance estimation for finite mixture models, J. Amer. Statist. Assoc. 91 (1996) 1716–1723.

[19] P. Deheuvels, Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité, C. R. Acad. Sci. Paris Sér. A 278 (1974) 1217–1220.

[20] L.P. Devroye, T.J. Wagner, The strong uniform consistency of kernel density estimates, J. Multivariate Anal. 5 (1980) 59–77.
[21] Y. Dodge, L_1-Statistical Procedures and Related Topics, Institute of Mathematical Statistics, Lecture Notes—Monograph Series, Hayward, CA, 1997.
[22] R.V. Foutz, On the unique consistent solution to the likelihood equation, J. Amer. Statist. Assoc. 72 (1977) 147–148.
[23] E. Giné, A. Guillou, Rates of strong uniform consistency for multivariate kernel density estimators, Ann. Inst. H. Poincaré 38 (2002) 907–921.
[24] L.M. Graves, Theory of Functions of Real Variables, McGraw-Hill, New York, 1946.
[25] F.R. Hampel, Contributions to the theory of robust estimation, Ph.D. Thesis, University of California, Berkeley, 1968.
[26] F.R. Hampel, The influence curve and its role in robust estimation, J. Amer. Statist. Assoc. 69 (1974) 383–393.
[27] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: the Approach Based on Influence Functions, Wiley, New York, 1986.
[28] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Statist. 35 (1964) 73–101.
[29] R.J. Karunamuni, K.L. Mehra, Improvements on strong uniform consistency of some known kernel estimates of a density and its derivatives, Statist. Probab. Lett. 9 (1990) 133–140.
[30] P.D.M. MacDonald, Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren, J. R. Statist. Soc. Ser. B 33 (1971) 326–329.
[31] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
[32] R.E. Quandt, J.B. Ramsey, Estimating mixtures of normal distributions and switching regressions, J. Amer. Statist. Assoc. (1978) 730–738.
[33] R. Ranga Rao, Relations between weak and uniform convergence of measures with applications, Ann. Math. Statist. 33 (1962) 659–680.
[34] B.W. Silverman, Weak and strong uniform consistency of the kernel estimate of a density and its derivatives, Ann. Statist. 6 (1978) 177–184 B.W. Silverman, Switching regressions, J. Amer. Statist. Assoc. (1978) 730–738.
[35] W.A. Woodward, W.C. Parr, W.R. Schucany, H. Lindsey, A comparison of minimum distance and maximum likelihood estimation of proportion, J. Amer. Statist. Assoc. 79 (1984) 590–598.